CONTENTS

1

ASSESSING TEACHING

A Changing Landscape

Amy E. Dayton

Assessing the teaching of writing is a process fraught with conflict. Despite a significant body of research pointing to the importance of multiple assessment measures and careful interpretation of the data, the evaluation of postsecondary teaching still relies heavily on a single measure of performance—the student ratings score—and interpretation of this score is often done in a hasty, haphazard fashion. Aside from student ratings, other data on teaching effectiveness tend to be collected in piecemeal fashion, without sufficient space for reflection and dialogue. When it comes to assessment, practical realities—including a lack of time, administrative resources, or knowledge about best practices—frequently trump our intentions to do a comprehensive job of evaluating classroom performance. Without clear guidelines for collecting and interpreting data, the outcome can be influenced by individual biases about what counts as evidence of good teaching. This collection offers new perspectives on that question of "what counts," pointing to ways that we can more effectively gather data about teaching and offering practical guidance for interpreting it. It also suggests ways we can improve our practice, mentor new teachers, foster dialogue about best practices, and make those practices more visible.

This book is for teachers who want to improve their practice, administrators and program directors who hire and train instructors, and faculty and staff in writing programs, centers for teaching and learning, and other instructional support units on college campuses. Although its primary audience is composition specialists, the collection offers practical suggestions and perspectives that apply to many contexts for postsecondary teaching. The tools presented in these chapters—mid-semester focus groups, student evaluations of instruction, classroom observations,

DOI: 10.7330/9780874219661.c001

teaching portfolios, and so on—are used across the disciplines, in many instructional settings. While some chapters focus on specific methods, others provide new frameworks for thinking about assessment. In her chapter on writing center(ed) assessment, for instance, Nichole Bennett describes a philosophy that could work for both writing programs and other sites for teacher training across campuses. This approach involves bringing teachers and tutors into the broader conversation about the program's missions and goals, and asking them to reflect on assessment data. By making assessment a broad, program-wide conversation, we invite stakeholders at every level to participate in setting goals and outcomes and gauging how well those outcomes have been met. The authors of chapters 6 and 7 argue for an ethos of transparency, suggesting a need to set clear standards for how materials might be read, to give teachers a sense of agency in deciding how to represent their work, and to share evidence of teaching quality with broader audiences while contextualizing the data for outside readers. These more inclusive, transparent models allow us to engage both internal and external audiences in more productive dialogue.

This collection arrives at a time when the public dialogue and political context for postsecondary teaching are particularly fraught. Challenges include a decline in state funding, public anxiety over the rising cost of college, concern about the value of a degree in today's lagging economy, and, to some extent, hostility toward college professors. An example of this hostility is found in Richard Arum and Josipa Roksa's recent book, *Academically Adrift*, which criticizes faculty for being more interested in their research and the advancement of their disciplines than in their students' progress or the well-being of their institutions—a trend that, in the authors' view, has contributed to an epidemic of "limited learning" on college campuses[1] (Arum and Roksa 2011, 10–11). (See Richard Haswell [2012] for a critique of their findings and methodology). At the state level, this picture of the self-interested, disengaged faculty member permeates our political rhetoric. The *Chronicle of Higher Education* reports that recent state election cycles have been dominated by efforts to curb faculty rights, including measures to limit salaries and collective bargaining rights, attacks on tenure and sabbaticals, and proposals to require college faculty to teach a minimum number of credit hours (Kelderman 2011). In a 2010 *Wall Street Journal* piece, "Putting a Price on Professors," Simon and Banchero (2010) point to some other developments. Texas state law now requires that public universities publicize departmental budgets, instructors' curriculum vitae, student ratings, and course syllabi, making all of this data accessible "within three

clicks" of the university's home page. At Texas A&M, university officials have gone even further, putting a controversial system in place to offer cash bonuses to faculty who earn the highest student ratings, and creating a public "profit and loss" statement on each faculty member that "[weighs] their annual salary against students taught, tuition generated, and research grants obtained" (Simon and Banchero 2010; see also Hamermesh 2010, Huckabee 2009, June 2010, and Mangan 2000).

This push to make college faculty more accountable—and to quantify their contributions—comes, ironically, at a time when tenured, sabbatical-eligible faculty members are dwindling in numbers, being replaced by part-time and non-tenure track teachers whose situations are often tenuous at best. A *New York Times* article reports that "only a quarter of the academic work force is tenured, or on track for tenure, down from more than a third in 1995" (Lewin 2013). The challenge facing many university writing programs, then, is not the task of fostering commitment to teaching among research-obsessed, tenured faculty members, but rather supporting teachers who are new to the profession—like graduate teaching assistants—or who are working without job security, a full-time income, or adequate professional resources (such as office space or support for professional development). Because first-year composition (FYC) is one of the few courses required for most students at public universities, and because personalized, process-based instruction requires low student-to-faculty ratios, university writing programs find themselves at the front lines of these labor issues in higher education.

Despite the challenging times, composition studies, as a field, has capitalized on the accountability movement and current zeal for assessment by taking a proactive stance, seeking meaningful ways to gather data about teaching and participate in large-scale evaluations of student learning. In the aftermath of the No Child Left Behind Act, the Spellings Commission on Higher Education, and initiatives such as the ones put in place in Texas, we recognize that developing thoughtful, context-sensitive assessments is the best insurance against having hasty, reductionist evaluations imposed upon our programs.[2] Many writing programs have either fully adopted the WPA Outcomes Statement on First-Year Composition (Council of Writing Program Administrators 2000), or have modified the statement to create local outcomes. Other programs are participating in large-scale, national assessments and making use of the data for local purposes. As Paine and his colleagues explain in chapter 11, the Council of Writing Program Administrators has teamed up with the consortium for the National Survey of Student

Engagement (NSSE) to create a writing component within that national assessment. In chapter 12, Deborah Goodburn and Amy Minter point to the ways that the trend toward "big data" can provide methods for analyzing trends and understanding patterns on our campuses and in our programs (they also acknowledge the need to use data mining in a responsible manner). These large-scale assessment projects have raised the visibility of our professional associations. More importantly, they have helped ensure that efforts to standardize outcomes or measure students' experiences with writing are informed by a solid understanding of composing processes and best practices for teaching.

While the national context for higher education has changed in recent years, the assessment landscape is also shifting. One way to gauge some of those changes is by considering the essays in this volume in relation to Christine Hult's 1994 text, *Evaluating Teachers of Writing*. On one hand, many of the central concerns of Hult's volume—the impact of the teaching-as-scholarship movement, the need to develop equitable practices to assess adjunct and graduate student teachers, and the overreliance on student surveys—are still important issues. On the other hand, the methods we use to assess teaching have evolved in ways that make them quite different from their predecessors. Take, for example, the practice of gathering mid-semester feedback from students. While Peter Elbow (1994), in the Hult volume, presents this method as an informal exchange between the students and the teacher, in chapter 5 of this volume Gerald Nelms explains how the small-group instructional diagnosis (SGID) method has formalized and systematized this practice, yielding more data and more reliable results. My point here is not that formal methods should always be privileged over informal, organic ones, but that with a range of methods at our disposal teachers have more choices about the kinds of feedback they would like to obtain.

Similarly, emerging technologies create new options for sharing our results. Electronic portfolios, teachers' homepages, professor rating websites, and other digital spaces now function not just to display data but also to foster conversation about their meaning. The dialogic nature of Web 2.0 technologies can make our assessments more open and transparent—but they also bring challenges for teachers who may not want to be visible in the way that technology allows (or compels) us to be. In chapters 7 and 10, Chris Anson and Amy Kimme Hea present contrasting perspectives on the tension between teachers' visibility and vulnerability online. While Anson urges writing programs and teachers to consider making assessment data more visible (by posting student opinion surveys online, for instance), Kimme Hea suggests ways that teachers can

monitor and manage their online presence, noting that today's teachers are being "written by the web" in ways we could not have predicted before the arrival of Web 2.0.

KEY TERMS

For readers who are new to the assessment landscape, the following section gives a brief overview of the key concepts that appear throughout this book. This section will also complicate these common terms, and will show how we might blur the boundaries between them in order to consider anew the potential, and the peril, of the approaches we choose.

Assessment

The term *assessment*, with its origins in the Latin phrase "to sit beside," suggests the possibilities inherent in formative, cooperative methods for training and mentoring writing instructors. Traditionally, composition scholarship, rooted in a humanist, progressive tradition that values the potential of the individual, has privileged that cooperative work of "sitting beside" our developing teachers over the sometimes necessary, but less pleasant, task of ranking, sorting, and judging them.

In recent years, writing assessment research has reached a kind of crossroads, with opposing visions of the work that we ought to be doing. On one hand, most scholars are deeply invested in empirical methods, drawing from the methodologies of our colleagues in educational measurement (Huot 2007; O'Neill, Moore, and Huot 2009; Wolcott and Legg 1998). These traditional approaches provide us with the means for gauging validity and reliability, as well as reading statistical results. On the other hand, an emerging body of work calls on composition scholars to take a more critical stance, and to interrogate the ideologies implicit in standardized assessments. Patricia Lynne (2004), for instance, rejects psychometric approaches entirely, advocating a rhetorically-based approach that eschews positivist assumptions, while Inoue and Poe (2012) urge us to consider "how unequal or unfair outcomes may be structured into our assessment technologies and the interpretations that we make from their outcomes" (6). That concern about the positivist assumptions and ideologies embedded in assessment work is not unique to scholars in the humanities, but is also the focus of an evolving conversation among scholars in the social sciences, including the field of educational measurement. In her influential essay, "Can There Be Validity without Reliability?" Pamela Moss (1994) argues that we cannot

make reliability judgments solely from statistical analyses of numerical data. Rather, they require an interpretive or "hermeneutic" approach involving "holistic, integrative" thinking "that [seeks] to understand the whole in light of its parts, that [privileges] readers who are most knowledgeable about the context in which the assessment occurs, and that [grounds] those interpretations not only in the . . . evidence available, but also in a rational debate among the community of interpreters" (7). In other words, assessment is, at least in part, a rhetorical practice, regardless of the disciplinary home of the person conducting the evaluation. When we assess, therefore, we must ask: Who are the stakeholders? Whom and what are we assessing? For what purposes? Who will be the ultimate audience? (Huot 2002; O'Neill, Moore, and Huot 2009). For this reason, most of the essays in this volume strike a balance between empirical and interpretive modes, without privileging one approach over the other.

Formative vs. Summative Assessment

Assessment scholars traditionally distinguish between formative and summative evaluation. Formative evaluation is "ongoing," designed to encourage improvement, while summative evaluation is "more fixed and 'retroactive,' bearing the connotation of finality in its sense of accountability" (Wolcott and Legg 1998, 4). Formative assessment is a tool to help teachers; it involves an element of self-evaluation that is best used in situations where instructors have the opportunity to reflect on the feedback, set goals for improvement, and implement the results in their classroom. Summative assessment, on the other hand, is done for external audiences, for the purpose of sorting, ranking, and making decisions about teachers—for example, when giving awards or making decisions about staffing, merit raises, contract renewals, and promotions.

In practice, the categories of formative and summative assessment are not clearly distinct from one another, nor should they be. Chris Anson argues in chapter 7 that summative evaluation should include some evidence of formative, or reflective, thinking about teaching. Moreover, when programs do not have the time and resources to offer both formative and summative evaluation (through multiple course observations, for instance), they tend not to make distinctions between them. It may be more productive, then, to use the term *instructive assessment*, as Brian Huot (2002) suggests. Instructive assessment shifts the focus to the teacher's growth and continuous improvement, even when making summative judgments. This stance reflects the growing consensus in

educational circles "[recognizing] the importance of holding all educational practices, including assessment, to rigorous standards that include the enhancement of teaching and learning" (18). This may be especially true for university writing programs. Considering the marginalized status of many of our teachers, it is critical that our assessments facilitate their continued improvement and professional development—and lead to some discussion about the resources our teachers need to be successful and the ways that programs and WPAs can provide better support.

Validity and Reliability

Almost all scholarly discussion of assessment begins with a review of the concepts of validity and reliability. In common parlance, validity—more specifically, *construct validity*—is thought of as the "truth" of an assessment, or the degree to which a test or tool measures what it purports to measure. When we say that a test or tool is "valid," we mean exactly that—it measures what it purports to measure. In assessment language, however, we tend not to make validity judgments about the tools themselves; rather, *validity* refers to the data produced by our instruments. That is, "tests [and other assessments] are not in and of themselves valid or invalid but rather the *results* are considered to be valid or invalid according to their intended use" (O'Neill, Moore and Huot 2009, 47, emphasis original). Determinations about validity include thoughtful interpretation of the data and careful construction of "a sound argument to support the interpretation and use of test scores from both theoretical and scholarly evidence" (O'Neill, Moore, and Huot 2009, 47). That evidence may include: the context of the evaluation, the process of administering it, the influence of external variables, and the consequences of the assessment (46–47). Thus, the emerging view of validity is that it is not "some pronouncement of approval but rather . . . an ongoing process of critical reflection" (Huot 2002, 51).[3] Another trend in our view of validity is the realization that we must attend to the ethical dimensions of the tools we have chosen, and that those aspects factor into our validity judgments. In chapter 3, I discuss the move toward *consequential validity*, the notion that our validity judgments must consider the results, whether intentional or unintentional, of how data (such as student opinion survey results) are interpreted.

In contrast, reliability refers to the *consistency* of results achieved over time from repeated use of the same instrument. For standardized assessment, reliability is generally thought of as a quantitative determination: for example, on large-scale writing tests, reliability is determined by the

scores assigned by trained readers. Yet, as Moss (1994) notes, for smaller-scale or non-standardized assessments (such as classroom observations or teaching portfolios), reliability is more difficult to establish through quantitative means—when it comes to qualitative evidence about teaching, reliability may be determined by an informed community using a process of thoughtful deliberation of shared norms (7).[4] For small-scale assessments (with the exception of student ratings), reliability can be seen as a set of values that includes "accuracy, dependability, stability, and consistency" (O'Neill, Moore, and Huot 2009, 52).

OVERVIEW OF THE BOOK

The book is divided into two sections. Each chapter in the first section focuses on a different method or tool for evaluation: heuristics for evaluating teaching, student evaluations, classroom observations, mid-semester focus group feedback, and teacher portfolios. In chapter 2, Meredith DeCosta and Duane Roen draw from Ernest Boyer's teaching-as-scholarship model to suggest a framework that attempts to capture the complexity and intellectual rigor that good teaching requires. Their chapter offers a set of heuristics that functions as both a theoretical guide and a generative tool for helping us to evaluate teaching and identify areas where more development may be needed. The other chapters in section I complicate and refine our understanding of well-established assessment methods, such as teaching portfolios, course observations, and student evaluations of instruction. Brian Jackson, for instance, suggests in chapter 4 that the course observation offers an opportunity for WPAs to practice macro-level teaching and see how programmatic goals and outcomes are understood at the classroom level. And chapters 5 and 6 suggest that by formalizing other methods, or implementing them via new technologies, we can transform them into tools that offer more data, or more opportunities for sharing data with various audiences.

The chapters in section II look beyond specific methods to unique contexts and emerging trends. As in the previous section, technology is an important component: emerging technologies (like e-portfolios) create new potential for assessment, but they also raise new challenges (chapters 7 and 10). Chapters 8 and 12 suggest ways that WPAs and other administrators can build a "shared language" for assessment among teachers, students, tutors, administrators, and other stakeholders, as well as make use of both "big" and "small" data. They offer guidance for teachers and programs on managing their online presence, both by monitoring feedback on external sites and, where appropriate,

by taking charge of their own data and making it accessible to the public (chapters 7 and 10). Sharing our results, Chris Anson suggests in chapter 7, allows writing teachers and programs to provide a rhetorical context in order to help the audience understand what student ratings or syllabi/course materials mean and how they are used. This section also offers pragmatic guidance. Cindy Moore, for instance, notes in chapter 9 that one of the biggest obstacles to good assessment is a lack of time and resources, and she offers suggestions for overcoming these challenges. Chapters 11 and 12 explore the trend toward large-scale assessment, suggesting ways that writing programs can use big data to better understand the dynamics of local programs.

ASSESSMENT AS RHETORICAL PRACTICE

Individually, the essays in this book address particular methods and models for assessment. Collectively, they present an argument for new ways of thinking about evaluating teaching. They respond to the public call to make teaching data more transparent and available for public discussion, and they suggest ways of contextualizing our assessments and using them to arrive at more nuanced understandings of what good teaching is. In making the case for new modes and models, they mimic the process of evaluation itself. When assessing their work, teachers and programs create narratives that illustrate what they have set out to do and show how they are working toward those goals and achieving results. Evaluating teaching, then, involves looking at a collection of evidence and analyzing and interpreting the argument it presents. When viewed in this light, the task of analyzing evidence, evaluating pedagogical approaches in their particular rhetorical context, and fostering a dialogue about best practices should present an appealing challenge for a group of scholars and teachers who are steeped in the scholarly tradition of critical interpretation, analysis, and thoughtful debate.

Notes

1. The actual findings of their study were far more modest than the broad claim the book suggests: their study found that college students made "modest gains" from the first semester of their freshman year to the second semester of their sophomore year.
2. See Brian Huot's (2007) "Consistently Inconsistent: Business and the Spellings Commission Report on Higher Education."
3. For a more thorough discussion of emerging views of validity and its types, see Messick (1989).

4. Moss (1994) notes that the danger of previous conceptions of validity and reliability—which insist upon quantification—is that they might lead us not to engage in good teaching and assessment practices merely because they are small scale, qualitative, and/or nonmeasurable (6).

References

Arum, Richard, and Josipa Roksa. 2011. *Academically Adrift: Limited Learning on College Campuses*. Chicago: University of Chicago Press.

Council of Writing Program Administrators. 2000. "WPA Outcomes Statement for First Year Composition." http://wpacouncil.org/positions/outcomes.html.

Elbow, Peter. 1994. "Making Better Use of Student Evaluations of Teachers." In *Evaluating Teachers of Writing*, ed. Christine Hult. Urbana, IL: NCTE.

Hamermesh, Daniel. 2010. "What's Your Econ 101 Professor Worth?" *The New York Times*, October 26.

Haswell, Richard. 2012. "Methodologically Adrift: Review of Arum and Roksa, *Academically Adrift*." *College Composition and Communication* 63 (3): 487–91.

Huckabee, Charles. 2009. "Professors Question Texas A&M's Plan to Award Bonuses on Basis of Student Evaluations." *The Chronicle of Higher Education*, January 11.

Hult, Christine, ed. 1994. *Evaluating Teachers of Writing*. Urbana, IL: NCTE.

Huot, Brian. 2002. *(Re)Articulating Writing Assessment for Teaching and Learning*. Logan: Utah State University Press.

Huot, Brian. 2007. "Consistently Inconsistent: Business and the Spellings Commission Report on Higher Education." *College English* 69 (5): 512–25.

Inoue, Asao, and Mya Poe. 2012. *Race and Writing Assessment*. New York: Lang.

June, Audrey Williams. 2010. "Texas A&M to Revise Controversial Faculty Rewards Based on Student Evaluations." *The Chronicle of Higher Education*, October 12.

Kelderman, Eric. 2011. "State Lawmakers Seek More Say over Colleges." *The Chronicle of Higher Education*, February 27.

Lewin, Tamar. 2013. "More College Adjuncts See Strength in Union Numbers." *The New York Times*, December 3.

Lynne, Patricia. 2004. *Coming to Terms: A Theory of Writing Assessment*. Logan: Utah State University Press.

Mangan, Katherine. 2000. "Bonus Pay Based on Student Evaluations Evokes Skepticism at Texas A&M." *The Chronicle of Higher Education*, January 20.

Messick, Samuel. 1989. "Validity." In R. L. Linn, *Educational Measurement*, 3rd ed. Old Tappan, NJ: Macmillan.

Minter, Deborah, and Amy Goodburn. 2002. *Composition, Pedagogy, and the Scholarship of Teaching*. Portsmouth, NH: Boynton/Cook.

Moss, Pamela. 1994. "Can There Be Validity without Reliability?" *Educational Researcher* 23 (2): 5–12. http://dx.doi.org/10.3102/0013189X023002005.

O'Neill, Peggy, Cindy Moore, and Brian Huot. 2009. *A Guide to College Writing Assessment*. Logan: Utah State University Press.

Simon, Stephanie, and Stephanie Banchero. 2010. "Putting a Price on Professors." *Wall Street Journal*, October 22.

Wolcott, Willa, and Sue M. Legg. 1998. *An Overview of Writing Assessment: Theory, Research, and Practice*. National Council of Teachers of English.